SJSU SAN JOSÉ STATE UNIVERSITY

**College of Science · Computer Science**

# Processing Big Data: Tools and Techniques Section 61

## CS 131

Summer 2024   3 Unit(s)   06/03/2024 to 08/09/2024   Modified 06/03/2024

## Contact Information

### Dr. Genya Ishigaki

**Email:** genya.ishigaki@sjsu.edu
**Office:** MH 215
**Phone:** (408) 924-5076
**Website:** https://sjsu-interconnect.github.io/ (https://sjsu-interconnect.github.io/)

#### Office Hours

Monday, Wednesday, 11:00 AM to 12:00 PM, MH 215

You don't need to make an appointment for these office hours. You can stop by my office.

## Course Description and Requisites

In-depth study of essential tools and techniques for processing big data over the UNIX operating system and/or other operating systems. On UNIX, it includes using grep, sed, awk, join, and programming advanced shell scripts for manipulating big data.

Prerequisite(s): CS 46B or BIOL 123B with a grade of "C-" or better. Allowed Declared Majors: Computer Science BS, Data Science BS, MS Bioinformatics (MS BI).

Letter Graded

## Classroom Protocols

### Communication with the instructor

Students are requested to use the Canvas message function to contact the instructor. Private messages sent to the instructor’s email address gets lost due to the large volume of emails received.

The instructor does not write messages after normal business hours, on weekends or holidays.

Reviewing code for the homework and technical trouble-shooting should be done during the office hours.

Never send your entire code for an assignment to the instructor. The instructor will not fix all the bugs in your code.

# Program Information

Diversity Statement - At SJSU, it is important to create a safe learning environment where we can explore, learn, and grow together. We strive to build a diverse, equitable, inclusive culture that values, encourages, and supports students from all backgrounds and experiences.

# Course Learning Outcomes (CLOs)

Upon successful completion of this course, students will be able to:

- Analyze, manipulate and process large-scale data with the UNIX/Linux command line and other operating systems.
- Develop shell scripts for use in data-intensive applications.
- Build data analysis pipelines, automate tasks, make analyses reproducible and shareable.
- Compare data analysis on the command line with use of graphical user interface and web-based tools.
- Solve big data challenges with the UNIX/Linux shell and command-line tools.
- Apply data science solutions to datasets from example domains, such as biology, business, and finance.
- Perform big data analysis efficiently, document and reproduce analysis, use cloud computing for data-intensive problems.

# Course Materials

Textbook:

- UNIX Command Line: A Complete Introduction. William Shotts Jr. [Download it (https://linuxcommand.org/tlcl.php) from the author's page]

Other good readings:

- Data Science at the Command Line, 2nd Edition. Jeroen Janssens, Publisher(s): O'Reilly Media, Inc. ISBN: 978149208791. [You can read it free through SJSU library (https://library.sjsu.edu/ebooks/safari-books-online-o-reilly).]
- Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python, 2nd Edition. Peter Bruce, Andrew Bruce, and Peter Gedeck, Publisher(s): O'Reilly Media, Inc. ISBN: 149207294X. [You can read it free through SJSU library (https://library.sjsu.edu/ebooks/safari-books-online-o-reilly).]
- Linux Journey. https://linuxjourney.com/

Technology:

- Practice of command-line operations will be done on IBM's computing cloud, Google Cloud and Amazon AWS. Instructions to subscribe for a free student account will be provided.
- Some assignments and worksheet tasks need to be submitted through Github. Details will be given in first assignment and worksheet instructions.

# Course Requirements and Assignments

### Exams

Three exams will be conducted during the regular class hours. A tentative schedule will be given in the course schedule below.

The exams will contain multiple choice questions, true/false and short answer questions. Exams are open book, open notes, and comprehensive. No make-up exams except in case of verifiable emergency circumstances.

### Homework assignments

There will be six assignments in total (a0 - a5). The evaluation of a2 and a5 includes oral presentations in class. Please check the tentative schedule below.

All assignment solutions that you submit must be completely your own work (i.e., your solution cannot be copied from another source, such as other students, the internet, etc.). While it is fine to discuss the worksheet/assignment solutions with other students, solutions submitted on Canvas should reflect your own efforts. Oral examination might be requested. All homework should be submitted on Canvas, not by e-mail.

### Data Analytics Course Hands-On Worksheets

Hands-on materials will be given through online Data Analytics course. The details will be explained in class.

# Grading Information

| Assignment | Grade Weight |
|---|---|
| Exam 1 | 15 % |
| Exam 2 | 15 % |
| Exam 3 | 15 % |
| Assignment 0 | 3 % |

| | |
|---|---|
| Assignment 1 | 10 % |
| Assignment 2 (+ oral presentation) | 10 % |
| Assignment 3 | 10 % |
| Assignment 4 | 10 % |
| Assignment 5 (+ oral presentation) | 10 % |
| Data Analytics Courses | 2 % |

### Extra-credits and Reworks

No extra-credit assignments or rework opportunities will be given.

### Late Submission

Late submissions within 24 hours will be deducted 10% of its final grade. Submissions over 24 hours late will have 20% grade deducted. Late submissions over 2 days will not be accepted.

### Missed Assignments or Exams

When students need to miss an assignment deadline or exam due to health conditions or any other emergency, it should be reported within ONE week after the due date.

### Final Grade Table

| Total Grade | Letter Grade |
|---|---|
| 97% and above | A plus |
| 92% to 96% | A |
| 90% to 91% | A minus |
| 87% to 89% | B plus |
| 82% to 86% | B |
| 80% to 81% | B minus |
| 77% to 79% | C plus |
| 72% to 76% | C |
| 70% to 71% | C minus |
| 67% to 69% | D plus |
| 62% to 66% | D |

| | |
|---|---|
| 60% to 61% | D minus |
| 59% and below | F |

# University Policies

Per University Policy S16-9 (PDF) (http://www.sjsu.edu/senate/docs/S16-9.pdf), relevant university policy concerning all courses, such as student responsibilities, academic integrity, accommodations, dropping and adding, consent for recording of class, etc. and available student services (e.g. learning assistance, counseling, and other resources) are listed on the Syllabus Information (https://www.sjsu.edu/curriculum/courses/syllabus-info.php) web page. Make sure to visit this page to review and be aware of these university policies and resources.

# Course Schedule

| Date | Topic 1 | Topic 2 | Note |
|---|---|---|---|
| 6/3 | Course Introduction | intro, git, ssh | |
| 6/5 | basic commands, man help, git | redirection, regex, vim | Due: a0 |
| 6/10 | redirection, regex, vim | Worksheet exercise 1 | |
| 6/12 | permission, process | monitoring, config .bashrc | |
| 6/17 | Worksheet exercise 2 | Exam Review 1 | Due: a1 |
| 6/19 | Juneteenth Day - Campus Closed | | |
| 6/24 | Exam 1 | shell scripting | |
| 6/26 | shell scripting | sed | |
| 7/1 | awk | awk | |
| 7/3 | a2 presentation | a2 presentation | Due: a2 Mini Project |
| 7/8 | Regression | Classification | |
| 7/10 | Command-line ML Exercise | Exam Review 2 | |
| 7/15 | Exam 2 | Command-line ML Exercise | |
| 7/17 | Python programming | pandas | Due: a3 |
| 7/22 | Exploratory Data Analysis | Visualization | |

| Date | Topic 1 | Topic 2 | Note |
|---|---|---|---|
| 7/24 | Visualization | make | |
| 7/29 | airflow | a5 discussion | Due: a4 |
| 7/31 | docker | Exam Review 3 | |
| 8/5 | Exam 3 | Exercise | |
| 8/7 | a5 presentation | a5 presentation | Due: a5 Mini Project, Data Analytics Courses |